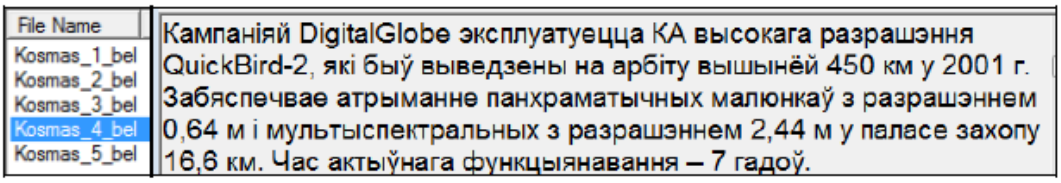
Ю.С. Гецэвіч, А.М. Скопінава (Мінск, НАН РБ)

ФАРМАЛІЗАЦЫЯ СТРУКТУР ЛІКАВЫХ ВЫРАЗАЎ З АДЗІНКАМІ ВЫМЯРЭННЯ Ў ТЭКСТАХ НА БЕЛАРУСКАЙ І РУСКАЙ МОВАХ

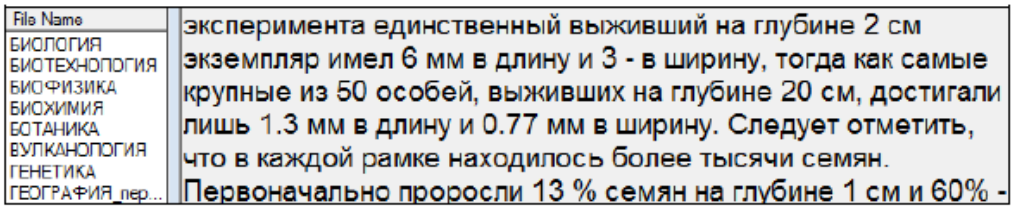
Распрацоўка сістэм сінтэзу маўлення па тэксце, пошуку і апрацоўкі інфармацыі, інтэлектуальных сістэм і г. д. непазбежна сутыкаецца з пытаннем пра спосабы распазнавання структураванай інфармацыі, напрыклад, у выглядзе літарна-сімвальных канструкцый. Колькасныя апісанні ўласцівыя агульнай навуковай карціне свету і, канешне, бытавой сферы жыцця. Такія апісанні прадстаўляюцца як колькасныя выразы ў спалучэнні з сістэмамі адзінак вымярэння (АВ), асабліва важнымі для метралогіі, матэматыкі, інфарматыкі, фізікі, тэорыі кадавання, прамысловасці, эканомікі, гандлю і інш. На XI міжнароднай канферэнцыі РІНТІ'2013 (г. Мінск, 15 лістапада 2012 г.) былі прадстаўленыя алгарытмы, якія знаходзяць колькасныя выразы з АВ і класіфікуюць іх па трох тыпах (СІ, вытворныя ад СІ, не СІ) з дакладнасцю 72 % [1]. У дадзеным дакладзе ставіцца **задача** распрацаваць сродкі фармалізацыі структур лікавых выразаў з АВ праз алгарытмы і лінгвістычныя рэсурсы для тэкстаў на беларускай і рускай мовах. **Складанасці** заключаюцца ў тым, што выразы з АВ маюць шырокую варыятыўнасць па напісанні і па ўтварэнні: *59 мА, 400 м/с, трыццаць пяць кілаграм, 200.5 кДж, 225 ккал, 35 руб., пяць галоўнаў паліва* і г. д.. Менавіта з-за гэтага практычна немагчыма пералічыць правілы лакалізацыі выразаў для ўсіх выпадкаў. Для спрашчэння гэтага працэсу патрэбна выкарыстоўваць прыстасаванні, якія дазваляюць зручна карэктаваць ужо распрацаваныя правілы і дадаваць новыя. У дадзеным даследаванні былі выкарыстаныя напрацоўкі пры пабудове беларускага і рускага модуляў міжнароднай камп'ютарна-лінгвістычнай праграмы NooJ [2], якая дазваляе будаваць алгарытмы пошуку складаных тэкставых фрагментаў у выглядзе візуальных выканальных графаў канчатковых аўтаматаў [3].

Перш за ўсё, патрэбна вызначыць заканамернасці ўтварэння колькасных выразаў з АВ. Для гэтага былі выкарыстаныя пабудаваныя з дапамогай NooJ беларуска- і рускамоўныя корпусы навукова-тэхнічных тэкстаў (мал. 1). Праз статыстычныя падлікі былі выбраныя тэксты 8 тэматык (ваеннае абсталяванне, геаграфія, гісторыя, касмічныя даследаванні, фізіка, мінералогія, транспарт і сувязь, батаніка) з найвялікшай колькасцю ўжыванняў лікаў. Затым у выглядзе рэгулярнага выраза эвалюцыйным шляхам была складзена формула для пошуку лікавых выразаў (мал. 2). Па меры аналізу правага кантэксту, формульны выраз

пастаянна паляпшаўся. У табліцы прадстаўленыя некаторыя вынікі спрацоўвання формулы для вышэй пералічаных тэкстаў.



а)



б)

Мал. 1. **Навукова-тэхнічны тэкставы корпус у фармаце NooJ для БМ *(а)* і РМ *(б)***

```
(от <NB> до <NB>)|(от <NB> до <NB><NB>)|(от <NB><NB> до <NB>)|(от <NB>,<NB> до <NB>,<NB>)|
(от <NB>,<NB> до <NB>)|(от <NB>.<NB> до <NB>.<NB>)|(от <NB> до <NB>,<NB>)|
(от <NB>–<NB> до <NB>–<NB>)|(от <NB>×<NB>-<NB> до <NB>×<NB>-<NB>)|
(от <NB>×<NB>–<NB> до <NB>×<NB>–<NB>)|(от <NB> до почти <NB>)|(<WF>|"~"|"=")(<NB>|(<NB><NB>)|
(<NB><NB><NB><NB><NB>)|(<NB><NB><NB><NB>)|(<NB><NB><NB>)|(<NB>×<NB>)|(<NB>,<NB>)|
(<NB>–<NB>)|(<NB>—<NB>)|(<NB>.<NB>)|(<NB>.<NB>.<NB>.<NB>)|(<NB>.<NB>)|(<NB>–<NB>,<NB>)|
(<NB>.<NB>.<NB>)|(<NB>,<NB>–<NB>,<NB>)|(<NB>,<NB>–<NB>)|(<NB>,<NB>.<NB>)|(<NB>.<NB>,<NB>)|
(<NB>"/"<NB>)|(<NB>":"<NB>)|(<NB>":"<NB>-<NB>)|(<NB>":"<NB>":"<NB>)|(<NB>,<NB>×<NB>–<NB>)|
(<NB>×<NB>–<NB>)|(<NB>´<NB>)|(<NB>,<NB>´<NB>,<NB>)|(<NB>´<NB>,<NB>)|(<NB>·<NB>)|
(<NB>,<NB>·<NB>–<NB>)| (<NB>.<NB>•<NB>)|(<NB>•<NB>–<NB>)|(<NB>.<NB>•<NB>–<NB>)|
(<NB>.<NB>·<NB>)|(<NB>,<NB>·<NB>)|(<NB>·<NB>–<NB>)|(">"<NB>)|(<NB>">>"<NB>))|
(<NB>|(<NB><NB>)|(<NB>,<NB>)|(<NB>–<NB>)|(<NB>—<NB>)|(<NB>-<NB>)|(<NB>.<NB>)|
(<NB>,<NB>–<NB>,<NB>)|(<NB>,<NB>–<NB>)|(<NB>-<NB>,<NB>)|(<NB>":"<NB>)|
(<NB>,<NB>×<NB>–<NB>)|(<NB>,<NB>·<NB>)|(<NB>·<NB>–<NB>)|(<NB>x<NB>)|(<NB>x<NB>')|
(<NB>x<NB>x<NB>)|(<NB>.<NB>x<NB>.<NB>)|(<NB>,<NB>×<NB>)|(<NB>¸<NB><NB>)|
```

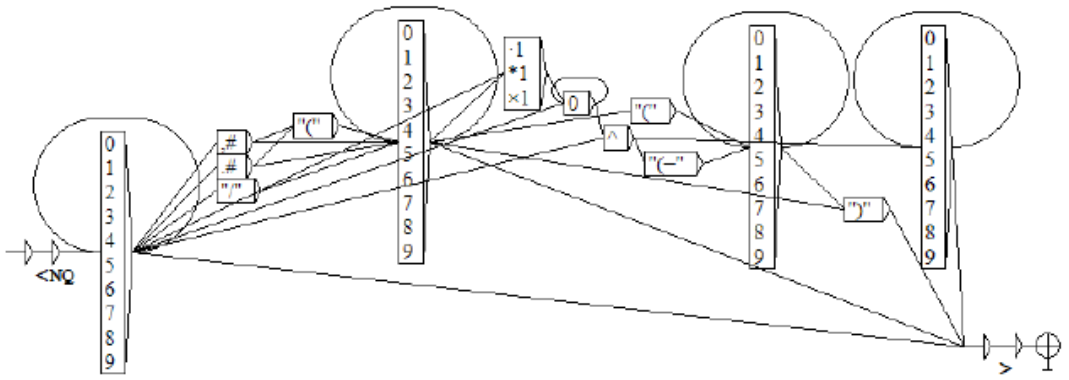
Мал. 2. **Рэгулярны выраз для пошуку структур лікавых выразаў**

Табліца

Фрагменты выніковага канкардансу па пошуку структур лікавых выразаў

Складнікі формулы	Прыклады знойдзеных выразаў
(от <NB> до <NB><NB>)	находятся в диапазоне от 4000 до 14 000 МГц и даже вариации) с периодами от 10 до 10 000 лет, сосредоточенными в
(от <NB>.<NB> до <NB>.<NB>)	диапазоне длин волн от 1.3 до 1.7 мкм. На подложке
(от <NB>–<NB> до <NB>–<NB>)	выемчатые. Их длина от 1–2 до 30–40 см. Самые длинные длиной волны l от 10–3 до 10–8 м. Этот диапазон
(от <NB>×<NB>-<NB> до <NB>×<NB>-<NB>)	удельным сопротивлением от 5×10-8 до 8×10-5 Ом×м. Композиционные
(от <NB>×<NB>–<NB> до <NB>×<NB>–<NB>)	в разных материалах: от 3×10–6 до 2×10–5 см. Магнитный поток
(от <NB> до почти <NB>)	током (при этом от 50 до почти 100 % его энергии превращается
(<WF>|"~"|"=")(<NB><NB>)	влений растениями страдают примерно 15 000 человек. Для дом ядерных взрывов суммарной силой 10 000 Мт в центральны»

Лікавыя структуры (лічбы, знакі ці лічбавыя выразы) — матэматычны аналаг колькаснага дэскрыптара, які часцей за ўсё стаіць перад АВ і разам з ёй утварае колькасны выраз. Колькасны дэскрыптар можа выражацца і лінгвістычна (пры дапамозе лічэбнікаў, колькасных займеннікаў, прыслоўяў і іх спалучэнняў), напрыклад: *тры молі, шмат градусаў, некалькі секунд* і да т.п.. На дадзеным этапе ажыццёўлены самастойны моўнезалежны алгарытм ідэнтыфікацыі колькасных дэскрыптараў (мал. 3), якія перададзеныя сродкамі матэматыкі і якія спрацоўваюць не толькі на простыя, дзесятковыя і дробавыя лічбы ў розных варыяцыях пісьмовага запісу, але і на лічбавыя выразы з экспаненцыяльнымі часткамі (мал. 4) [4].



Мал. 3. **Моўнезалежны алгарытм для ідэнтыфікацыі колькасна-лічбавых дэскрыптараў**

Before	Seq.
цела масай	102
прычэпа; · -	13,5
нага святла:	2,61·10^(-1
нім складае	5·10^(-5)

а)

Before	Seq.
не более	110
осями; · -	18,75
ставляет	5·10^(–5)
ота 0°) —	3,1·10^(–5)
ы: около	6·10^13

б)

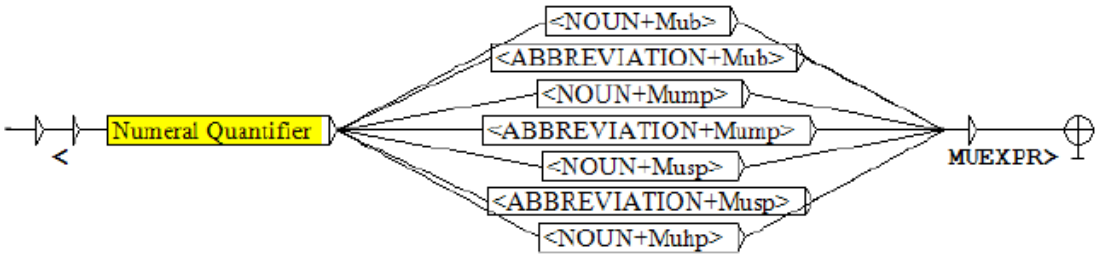
Before	Seq.
equal to	6.24150974×10
equal to	2.3901×10^(−4)
exactly	0.0254
exactly	453.59237
valent to	1/100

в)

Мал. 4. **Прыклады вынікаў пошуку колькасна-лічбавых дэскрыптараў для *(а)* беларускай, *(б)* рускай, *(в)* англійскай моў.**

Аналіз правага (адносна знойдзеных структур лікавых выразаў) кантэксту дазволіў класіфікаваць АВ па словаўтваральным прыкметам: з цэлай асновай і без прэфікса (*метр, Герц, Ом*); з цэлай асновай і з цэлым прэфіксам (*нанафарады, міліампер*); з цэлай асновай і са скарочаным прэфіксам (*кБайт*); са скарочанай асновай і без прэфікса (*Дж, га, Па*); са скарочанай асновай і са скарочаным прэфіксам (*км, дл, гПа*). Гэтая класіфікацыя была выкарыстана як галоўны прынцып пры распрацоўцы шматкампанентнага комплексу для ідэнтыфікацыі колькасных выразаў з АВ (мал. 5), які складаецца са слоўніка асноў (базавых АВ без прыставак і скарачэнняў), марфалагічных і сінтаксічных кампанентаў. У выніку яго працы АВ могуць атрымаць адзін з наступных маркераў: **Mump** (з кратным прэфіксам, **Musp** (з дольным прэфіксам), **Muhp** (з некалькімі прэфіксамі (напрыклад, *мікрамегафарад*)). Такім спосабам утвараць АВ не прадугледжана ў CI, таму дадзеныя словы трэба абазначыць у тэксце, каб пазней вывесці спіс памылкова ўтвораных АВ. У агульным кожнаму колькаснаму выразу з АВ

прысвойваецца маркер **<MUEXPR>**, паводле яго будуюцца выніковыя мэтавыя канкардансы (мал. 6).



Мал. 5. **Алгарытм для пошуку і ідэнтыфікацыі структур лікавых выразаў паводле словаўтваральных прыкмет**

Before	Seq.	A
ам (0,001 кг).	31 мкТл	(
уецца зблізку	2,4 мЗв	)
апору з сілай	9.81 Н	.
адамі (пішуць	60 000 пф	,
а з радыусам	1 сантыметр	,
о проста Mb).	1 мегабіт	:
ў вакууме за (	1 / 299 792 458) секунды	.

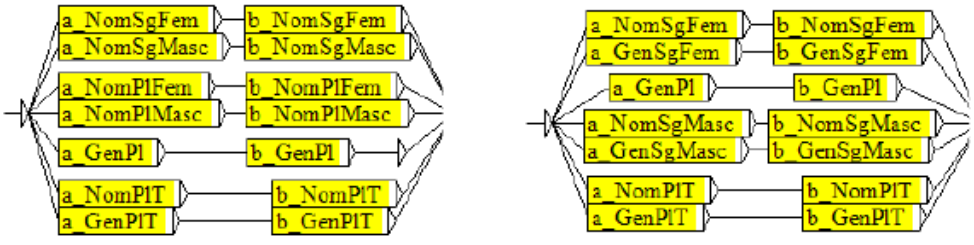
а)

Before	Seq.	Aft
евышал	1 мА	. Ha
кфайл в	100 килобайт	»).
от 1 до	100 МОм	, чт
равольт	13 йоттайоктограммов	Ka
бит/с) и	137,4 МГц	(м
массой	1383,95 каратов	. I

б)

Мал. 6. **Фрагменты вынікаў ідэнтыфікацыі колькасных выразаў з АВ паводле словаўтваральных прыкмет на матэрыяле беларуска-** ***(а)*** **і руска-** ***(б)*** **-моўных тэкстаў**

Акрамя ідэнтыфікацыі, у дачыненні да сінтэзатараў маўлення па тэксце важна распрацаваць метады **генеравання** арфаграфічных слоў па складанаструктураваных выразах. Напрыклад, выраз *7000 м* павінен “разгарнуцца” ў *сем тысяч метраў*. З гэтай мэтай у адносінах да колькасных выразаў з АВ былі распрацаваныя разгалінаваныя алгарытмічныя комплексы, якія складаюцца з 41 графа (падграфа), для беларускай і рускай моў (мал. 7). Сам алгарытм утрымоўвае графы 2 тыпаў: з назвамі на **a_** генеруюць лікі ад 0 да 999 999 999 999; з назвамі на **b_** прызначаюцца для токенаў, якія абазначаюць АВ. Колькасны выраз з АВ пасля ўваходу трапляе на выхад па адным з 7 шляхоў у залежнасці ад асаблівасцяў скланення назоўнікаў (з улікам катэгорыі роду) пасля лічбаў.



а) б)

Мал. 7. **Галоўны граф алгарытма генерацыі арфаграфічных слоў з колькасных выразаў з АВ для** ***(а)*** **беларускай і** ***(б)*** **рускай моў**

Напрыклад, пасля лічбы **1** (уключаючы лічбы з апошняй **1**) назоўнікі прымаюць форму назоўнага склону адзіночнага ліку (**NomSg**), акрамя так званых назоўнікаў pluralia tantum (напрыклад, *суткі*), якія набываюць канчатак назоўнага склону множнага ліку (**NomPlT**). Колькасныя выразы з назоўнікамі

мужчынскага (**Masc**) і жаночага (**Fem**) родаў апынуцца ў адпаведных падграфах **b_NomSgMasc** або **b_NomSgFem**. Варта адзначыць, што калі лічба заканчваецца на **2**, **3**, **4**, у беларускай мове ад назоўніка патрабуецца форма назоўнага склону множнага ліку (**NomPl**), у той час як у рускай — форма роднага склону адзіночнага ліку (**GenSg**). Калі ж справа даходзіць да астатніх лічбаў, то ў абедзвюх мовах патрабуецца канчатак жаночага роду множнага ліку (**GenPl**). Вынікі працы алгарытму адлюстраваныя на мал. 8.

700001г/семсот тысяч адна гадзіна
0 с/нуль секунд
777'700т/семсот семдзесят сем тысяч семсот тон
888'808хв/восемсот восемдзесят восем тысяч восемсот восем хвілін
2220020 хвіл/два мільёны дзвесце дваццаць тысяч дваццаць хвілін
444'014моль/чатырыста сорак чатыры тысячы чатырнаццаць моляў

а)

6661с/шесть тысяч шестьсот шестьдесят одна секунда
77700 т/семьдесят семь тысяч семьсот тонн
800009кд/восемьсот тысяч девять кандел
120202 мин/сто двадцать тысяч двести две минуты
8600км/ч/восемь тысяч шестьсот километров в час
903 м/с/девятьсот три метра в секунду

б)

Мал. 8. **Фрагменты вынікаў працы алгарытма генерацыі арфаграфічных слоў з колькасных выразаў з АВ для *(а)* беларускай і *(б)* рускай моў**

У **выніку** была пастаўлена і вырашана задача фармалізацыі структур колькасных выразаў з АВ у тэкстах на беларускай і рускай мовах навукова-тэхнічнай тэматыкі праз рэалізацыю алгарытмаў ідэнтыфікацыі і генеравання ў форме канчатковых аўтаматаў праз праграму NooJ. Канчатковыя аўтаматы наглядна паказваюць працу алгарытмаў і абазначаюць спосаб іх далейшага маштабавання і папаўнення лінгвістычнымі рэсурсамі для павышэння паказчыкаў паўнаты пры ацэнцы якасці. У будучым гэта **плануецца** зрабіць праз змяншэнне колькасці памылак пры ідэнтыфікацыі шматзначных выразаў (*г* для *год*, *грам*, *гадзіна*); распрацоўку алгарытмаў для колькасных дэскрыптараў, выражаных лінгвістычнымі сродкамі; папаўненне базы АВ менш ужывальнымі велічынямі.

ЛІТАРАТУРА

1. Гецэвіч, Ю.С. Ідэнтыфікацыя выразаў з адзінкамі вымярэння ў навукова-тэхнічных і прававых тэкстах на беларускай і рускай мовах / Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2012) : даклады XI Міжнар. канф. (Мінск, 15 лістапада 2012 г.). — Мінск, 2012. — С. 260–265.

2. Hetsevich, Y. Overview of Belarusian and Russian dictionaries and their adaptation for NooJ / Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 Intern. Conf. / eds. K. Vučković, B. Bekavac, M. Silberztein. — Newcastle : Cambridge Scholars Publishing, 2012. — P. 29–40.

3. Лінгвістычны працэсар NooJ [Электронны рэсурс]. — 2002. — Рэжым доступу: http://www.nooj4nlp.net/pages/nooj.html. — Дата доступу: 01.03.2013.

4. Гецэвіч, Ю.С. Кампаненты ідэнтыфікацыі колькасных выразаў з адзінкамі вымярэння ў тэкстах на беларускай і рускай мовах / Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS–2013) : материалы III Междунар. науч.-техн. конф. (Минск, 21–23 февраля 2013 года) / редкол.: В.В. Голенков (отв. ред.) [и др.]. — Минск, 2013. — С. 319–328.